DOCUMENT RESUME

ED 100 137 FL 005 200

AUTHOR	Barton, Ian J.; And Others
TITLE	Variable-Length Character String Analyses of Three Data-Bases, and their Application for File Compression.
PUB DATE	Apr 73
NOTE	14p.; Paper presented at the ASLIB Annual Conference, University of Durham, April, 1973
AVAILABLE FROM	ASLIB, 3 Belgrave Square, London SW1, England (6 pounds, 75 pence, for proceedings of conference)
EDRS PRICE	MF-$0.75 HC Not Available from EDRS. PLUS POSTAGE
DESCRIPTORS	*Computational Linguistics; *Computer Programs; Content Analysis; *Data Bases; Information Science; *Information Storage; *Programing Languages; Statistical Analysis; Word Frequency

ABSTRACT

A novel text analysis and characterization method involves the generation from text samples of sets of variable-length character strings. These sets are intermediate in number between the character set and the total number of words in a data base; their distribution is less disparate than those of either characters or words. The size of the sets of character strings (key-sets) can be varied arbitrarily by changing parameters. The characteristics of three scientific data bases (two disciplinary, one interdisciplinary) are compared in terms of key-sets of different sizes. Application of the key-sets for file compression, using a variable to fixed-length coding strategy, is discussed. (Author)

ED 100137

Variable-length character string analyses of three data-bases, and their application for file compression.

Ian J. Barton, Michael F. Lynch, J. Howard Petrie and Michael J. Snell.

Postgraduate School of Librarianship and Information Science
University of Sheffield

FL 005 200

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY
ASLIB
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

Abstract

A novel text analysis and characterisation method involves the generation from text samples of sets of variable-length character strings. These sets are intermediate in number between the character set and the total number of words in a data-base; their distribution is less disparate than those of either characters or words. The size of the sets of character strings (key-sets) can be varied arbitrarily by changing parameters.

The characteristics of three scientific data-bases (two disciplinary, one interdisciplinary) are compared in terms of key-sets of different sizes. Application of the key-sets for file compression, using a variable to fixed-length coding strategy, is discussed.

Introduction.

Shannon[1] tells us that the set of symbols ideal for economy in mechanical storage and transmission of information is one in which the symbols are equiprobable. In that case, the value for the entropy, as given by the expression

$$-H = \sum_{i=1}^{i} p_i \log_2 p_i$$

reaches a maximum. This value is the binary logarithm of the total number of symbols in the set, i.e., their variety. Since, in mechanical systems, symbols are most conveniently represented by fixed-length binary patterns, it is natural to consider symbol sets which are equal in number to integral powers of two. A series of such ideal sets can be represented as in Figure 1.

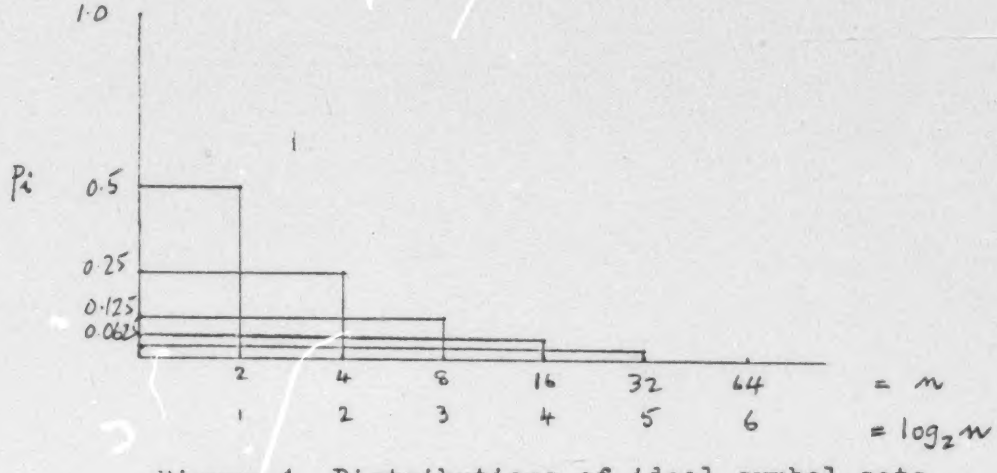

Figure 1. Distributions of ideal symbol sets.

The two dominant features are that the distribution is rectangular, and that the value of the entropy is determined by the variety of symbols, however defined.

Symbol sets with such ideal distributions are most uncommon in natural circumstances; much more typical is a hyperbolic distribution[2], such as that displayed by the characters of the titles of articles included in Chemical Titles, as shown in Figure 2.

The most common approach to converting a hyperbolic distribution to a rectangular one is exemplified by schemes

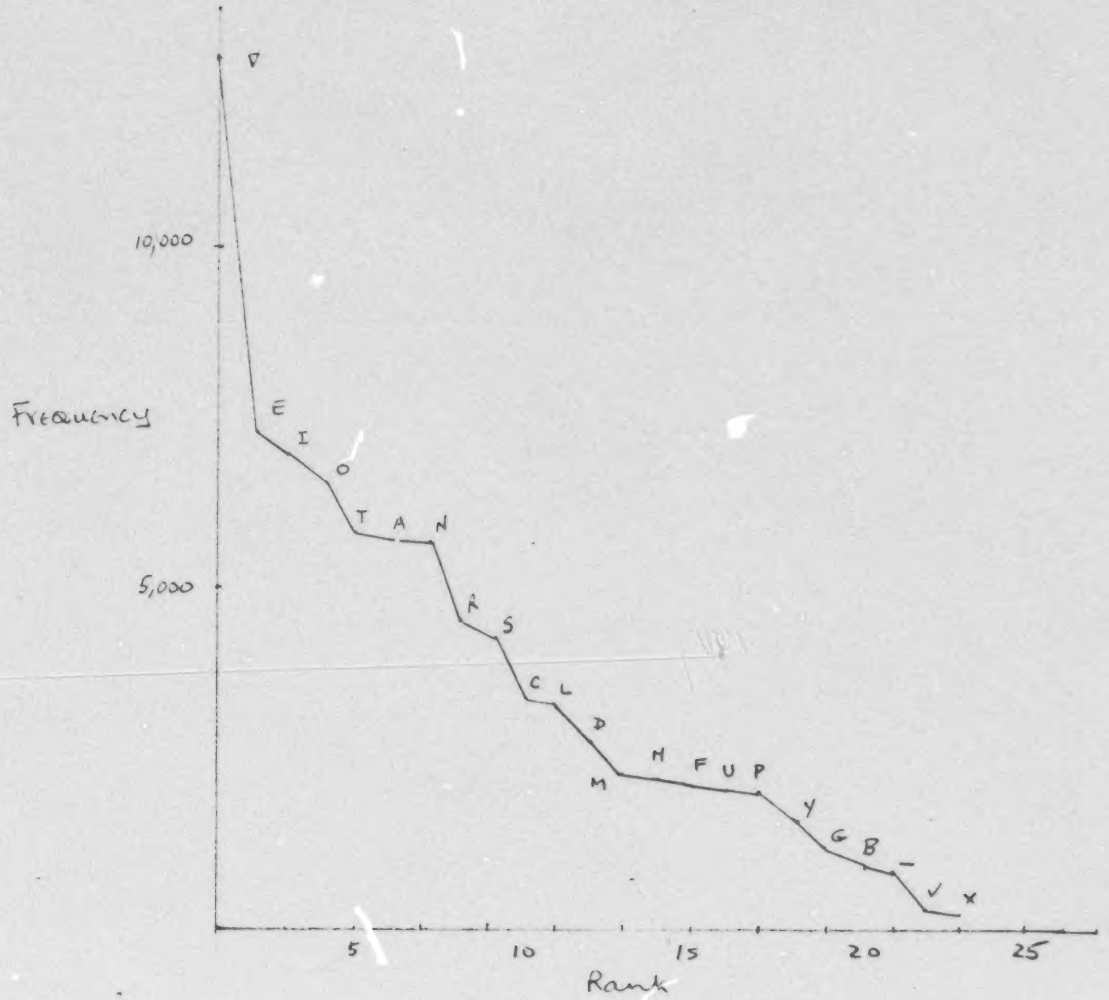

Figure 2. Distribution of characters in 1000 titles from Chemical Titles.

introduced by Shannon, Fano[3] and by Huffman[4]. These involve taking a fixed-length segment of text, and representing it by means of a variable-length code, the length of the code being inversely related to the frequency of the symbol. This can be shown - if notionally - by the diagram of Figure 3. It involves a fixed-to-variable length transformation, which is merely one of three possible strategies, the others being variable-to-fixed length, and variable-to-variable length. The second is known in the context of run-length coding, a method used for compression of graphic data[5], but has otherwise not attracted

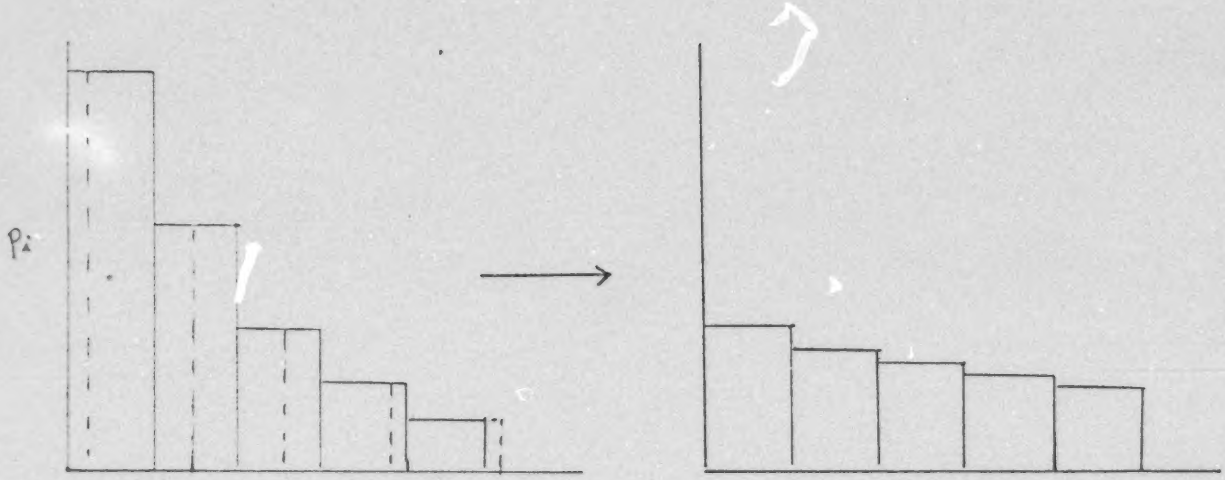

Figure 3. Notional mapping of hyperbolic onto rectangular distribution by means of fixed-to-variable length transformation.

much study, although Walker[6] has used it to compress Christian names taken from early English parish registers.

If we consider a hyperbolic symbol distribution, the disparate frequencies can be reduced by considering uniform aggregates of the symbols. Thus, if we count digrams, i.e., character pairs, instead of symbols, we can reduce the greatest frequency by perhaps an order of magnitude. The cost of this is an increase in the variety of the new symbols considered, again by something like an order of magnitude. As we consider longer uniform character strings, or fixed-length n-grams, we constantly reduce the range of frequencies, but always with an accompanying increase in variety, as illustrated for INSPEC titles in Figure 4 for n = 1, 4 and 8.

Variable-length character strings.

Returning to the variable-to-fixed length compression approach, let us consider a simple method of ironing out unevenness in distribution without the great increase in

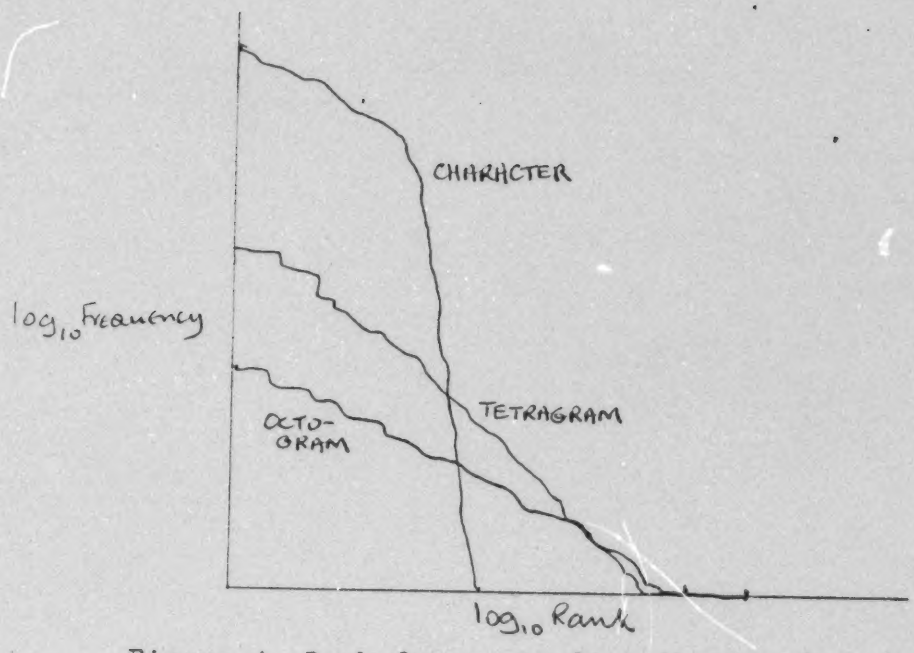

Figure 4. Rank frequency distribution for character strings of lengths 1, 4 and 8.

variety caused by taking uniform segments of text.

We first produce ranked lists of characters, digrams, trigrams, etc., for a substantial sample of text, as illustrated in Table 1 for Chemical Titles. By adding the most frequent digram to the original symbol set, we increase the variety by one, but reduce the frequencies of two of the most frequent characters quite substantially. We continue this by adding further digrams until we reach that digram with a frequency equal to or just below that of the most frequent trigram. We add this trigram in turn to the set. We continue the procedure, adding further n-grams of any length as their frequency equals that currently being considered, until the total number of "symbols" in the new set equals some power of two, e.g., 256.

This now constitutes what we call a key-set, composed of variable-length strings, or keys. The majority of these are short, digrams or trigrams. To each is assigned a numeric code, which can be represented by 8 bits for a key-set size of 256. Obviously, the process can be continued by continual addition

of further n-grams, so that the key-set can be enlarged to any desired level. Table 1 illustrates the ranked n-grams for a sample of 1000 titles from Chemical Titles.

n-gram length =	1		2		3		4		5	
	▽	12793	IN	1578	OF▽	1177	▽OF▽	1067	TION▽	593
	E	7440	E▽	1578	▽OF	1177	TION	843	ATION	497
	I	6902	TI	1529	ION	910	ION▽	640	▽AND▽	446
	O	6659	ON	1526	ON▽	843	ATIO	497	ION▽O	371
	T	5882	N▽	1514	TIO	805	AND▽	453	N▽OF▽	369
	A	5769	▽O	1496	ATI	595	▽AND	447	ON▽OF	363
	N	5733	▽A	1441	▽AN	582	▽THE	421	S▽OF▽	336

Table 1. Frequency-ranked n-grams from 1000 titles from Chemical Titles.

Data-Compression.

We now apply the key-set to text in the following manner. We take the initial characters, select the longest key available which matches it exactly, and substitute its code for the string. Starting from the next character not included in the first string, we repeat the process until the end of the text is reached. Figure 5 illustrates the process.

VARIABLE-LENGTH STRINGS

V B LE TH N
AR LE N ▽ST G
IA - G RI S

Figure 5. Encodement of text by variable-length character strings.

Obviously, the key-set must contain all characters in the text to be processed, no matter how infrequently some may appear.

When a code for a single character is used there may be an overall loss (depending on the fixed-length character code employed). When a longer n-gram is encoded, the number of bits required for storage is reduced, the advantage being the number of bits saved multiplied by the number of occasions on which the n-gram is used.

The method of selecting a key-set we have just outlined is a simple one, and ignores the fact that certain of the smaller keys are wholly contained within longer ones, and seldom if ever assigned. By eliminating these, and adding further n-grams from the candidate list, performance can be appreciably increased. We have already described other methods of generating key-sets by suitable programs[7,8]. Simple though the above method is, its performance is comparable with key-sets produced automatically.

We have now determined compression ratios obtainable with automatically produced key-sets on titles from three different data-bases, at two key-set sizes, 256 and 512. The composition of a typical key-set with 256 n-grams is shown in Table 2, while Table 3 shows the compression ratios obtained with these key-sets. The figures represent the reduction in the numbers of bits required for storage, based on a 6-bit character code (ICL 1900 Series computer). If the character code were an 8-bit code, the advantage gained would be correspondingly greater, reaching approximately 50% with the 256 key-set. This would presume use of a single-case character set, and of a shift-code if a multiple-case alphabet were used. It is worth noting that Snyderman and Hunt[9] and Schieber and Thomas[10] by adding digrams to the basic character code of IBM 360 machines have achieved compression ratios of 35% and 43·5%, while Byrne and Mullaney[11] have attained a ratio of 44%

n=						
9	ATION▽OF▽					
8	▽OF▽THE▽	TION▽OF▽				
7	ION▽OF▽	OF▽THE▽				
6	ATION▽	F▽THE▽	N▽THE▽			
	ON▽OF▽	TIONS▽				
5	▽AND▽	▽FOR▽	▽THE▽			
	CTION	IONS▽	N▽OF▽			
	S▽OF▽	TION▽				
4	▽IN▽	▽OF▽	▽ON▽	AND▽		
	ECTR	FOR▽	ING▽	ION▽		
	MENT	ONS▽	S▽IN			
3	▽A	▽CO	▽DE	▽DI	▽IN	▽ME
	etc.	(TOTAL = 42)				
2	▽A	▽B	▽C	▽E	etc.	
	(TOTAL = 130)					
1	55 CHARACTERS					

Table 2. Composition of key-set of 256 keys from INSPEC titles.

COMPRESSION RATIOS	
256 KEY-SET	
CT	33·9%
ASCA	31·7%
INSPEC	33.9%
512 KEY-SET	
CT	37·5%
ASCA	35·1%
INSPEC	37·6%

Table 3. Compression ratios obtained with key-sets of size 256 and 512 keys, (ratios based on a 6 bit character coding).

based on an 8-bit code, also using an n-gram method.

It is interesting to examine the extent to which the mapping of a hyperbolic distribution of characters onto a rectangular distribution has been achieved by this procedure. Figure 6 shows the shape of the rank-frequency distribution curve for the INSPEC 256 key-set applied to titles, plotted on a log/log scale. The entropy value was calculated and found

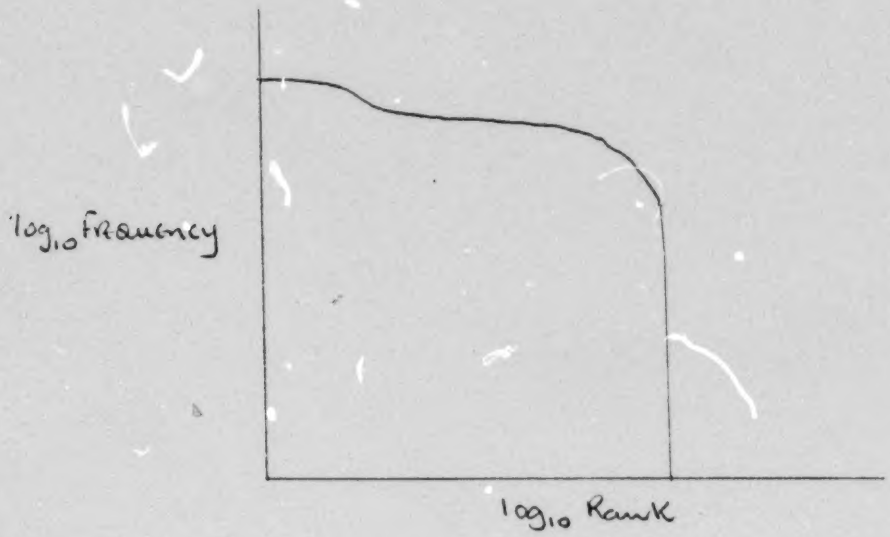

Figure 6. Rank-frequency curve (log/log) of distribution of n-gram keys.

to be 7·58, indicating that little further improvement of performance can be expected with this key-set size, although greater degrees of compression might well be obtained by using a variable-variable length strategy, the third mentioned above.

Comparison of data-bases.

Having earlier determined that key-sets produced from one data-base over a period of three years were substantially stable, we were interested in determining what similarities or differences existed between different data-bases. Those we chose were INSPEC, Chemical Titles and ASCA, representing two disciplinary data-bases and one interdisciplinary. Using key-sets containing 256 keys, 191 keys were found to be common

to all three, while pair-wise comparisons showed even greater similarities,i.e., 209 keys common to the sets from CT and ASCA, 210 common to CT and INSPEC, and 213 common to ASCA and INSPEC. Further confirmation of this similarity was obtained by using a key-set produced from one data-base for compression of another. As Table 4 illustrates, only slight reductions in

	256 Keys	
CT with INSPEC	31·9%	(33·9%)
INSPEC with ASCA	32·5%	(33·9%)
ASCA with CT	31·1%	(31·7%)
	512 Keys	
CT with INSPEC	34·3%	(37·5%)
INSPEC with ASCA	35·8%	(37·6%)
ASCA with CT	33·6%	(35·1%)

Table 4. Compression ratios using key-sets derived from another data-base.

the compression ratios were observed, indicating that as far as titles are concerned, the statistical microstructures of the data-bases are very similar indeed. This is in spite of substantial dissimilarities in the vocabularies of the data-bases; Table 5 gives the ranks of most frequent words in a sample of each data-base. (The word THE is automatically removed from titles in the ASCA file.) It is noticeable that discipline-oriented content words appear at higher ranks in INSPEC and CT than in ASCA.

Summary.

The work we have described represents an extension of the strategies available for data-compression, with potentially useful applications. It also provides an information-theoretic

ASCA		CT		INSPEC	
OF	1242	OF	1349	OF	1219
IN	523	AND	505	THE	782
AND	467	THE	496	IN	459
ON	205	IN	495	AND	450
FOR	136	ON	170	A	318
WITH	111	A	135	FOR	235
BY	106	BY	112	ON	204
A	102	WITH	101	WITH	105
TO	99	EFFECT	97	BY	89
FROM	61	FROM	92	TO	89
STUDIES	51	FOR	83	AN	79
STUDY	47	TO	69	FROM	60
EFFECT	46	DI	66	SCATTERING	69
EFFECTS	42	ACID	57	SYSTEM	56
NEW	39	MAGNETIC	54	ENERGY	55
AT	33	PROPERTIES	53	THEORY	54
BETWEEN	31	SYNTHESIS	47	AT	51
SOME	31	EFFECTS	42	SYSTEMS	46
DURING	30	NUCLEAR	39	MODEL	45
METHOD	30	OXIDE	39	STUDY	44
AN	29	STRUCTURE	39	CONTROL	42
GROWTH	27	DETERMINATION	38	METHOD	42
AS	26	METHYL	36	FIELD	40
PROPERTIES	26	SYSTEM	36	ANALYSIS	37
POLYMERIZATION	24	CHLORIDE	34	MAGNETIC	35
KINETICS	23	ACTIVITY	33	ELECTRON	34
REACTION	23	RESONANCE	33	PROPERTIES	34
USE	22	RAT	32	EFFECT	33
INFLUENCE	21	REACTION	30	RESONANCE	31

Table 5. Word rankings for samples of three data-bases.

model which we believe can have considerable significance in the context of computer-based retrieval systems, on which we hope to report shortly.

Acknowledgements.

We are grateful to OSTI for funds in support of this work, and to INSPEC, UKCIS and ISI for the provision of data-base samples.

References.

1. C. E. Shannon, The mathematical theory of communication, Bell Syst. Tech. J., 27, 398-403 (1948).

2. R. A. Fairthorne, Empirical frequency distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction, J. Doc., 25, 319-343 (1969).

3. R. M. Fano, Transmission of Information, New York and London, M.I.T. Press and John Wiley, 1961.

4. D.A. Huffman, A method for the construction of minimum redundancy codes, Proc. IRE, 40, 1098-1110 (1952).

5. S. W. Golomb, Run-length encodings, IEEE Trans. Inf. Theor., IT-12, 399-401 (1966).

6. V. D. Walker, Compaction of names by x-grams, in Am. Soc. Inf. Sci., Proc. 32nd Ann. Mtg., San Francisco, Oct. 1969, Vol. 6, Westport, Conn., and London, Greenwood Pub. Co., pp. 129-35.

7. M. F. Lynch, J. H. Petrie and M. J. Snell, The microstructure of titles in the INSPEC data-base, Inf. Stor. Retr. (in press)

8. A. C. Clare, E. M. Cook and M. F. Lynch, The identification of variable-length, equifrequent character strings in a natural language data-base, Comput. J., 15, 259-62 (1972).

9. M. Snyderman and B. Hunt, The myriad virtues of text compaction, Datamation, 1970, 36-40 (Dec. 1).

10. W. D. Schieber and G. W. Thomas, An algorithm for compaction of alphanumeric data, J. Lib. Aut., 4, 198-206 (1971).

11. J. G. Byrne and A. Mullaney, pers. comm.